Computational Social Science

What it is and **What it matters**

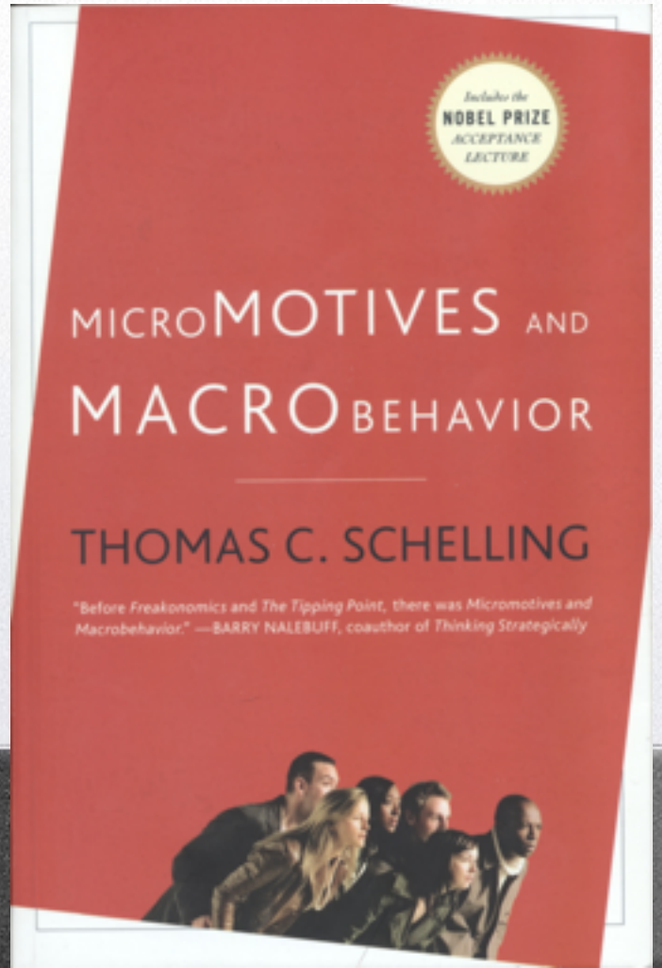
What is Computational Social Science?

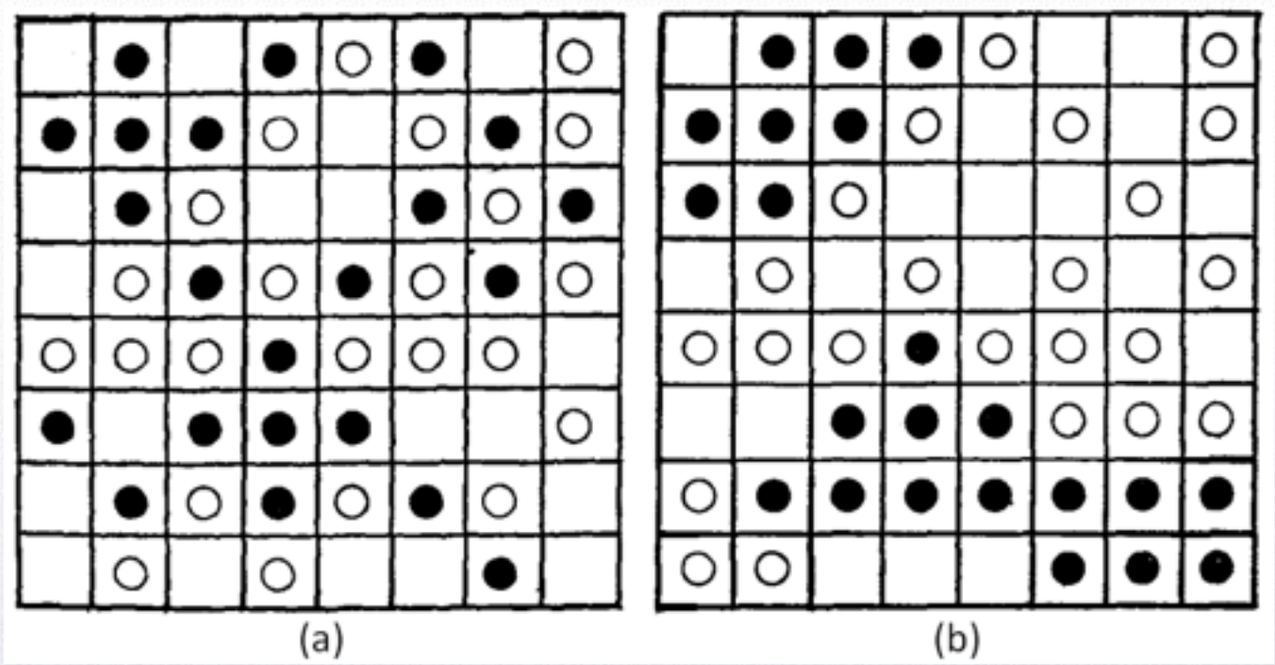
Using computers to generate data, discover patterns or generate and test explanations that you could not have without them.

…implies a shift in computationally enabled research designs, methods and theoretical standards

Before 2000

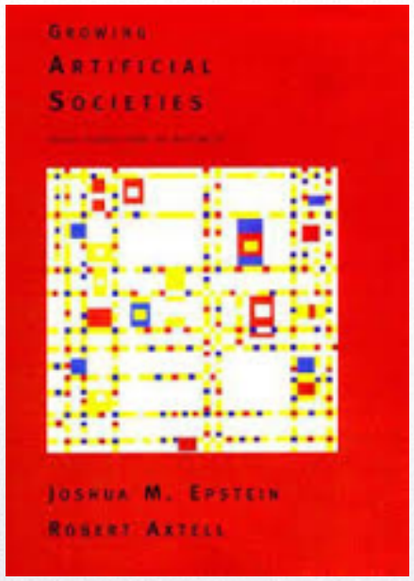
Computing consequences of theoretical assumptions…

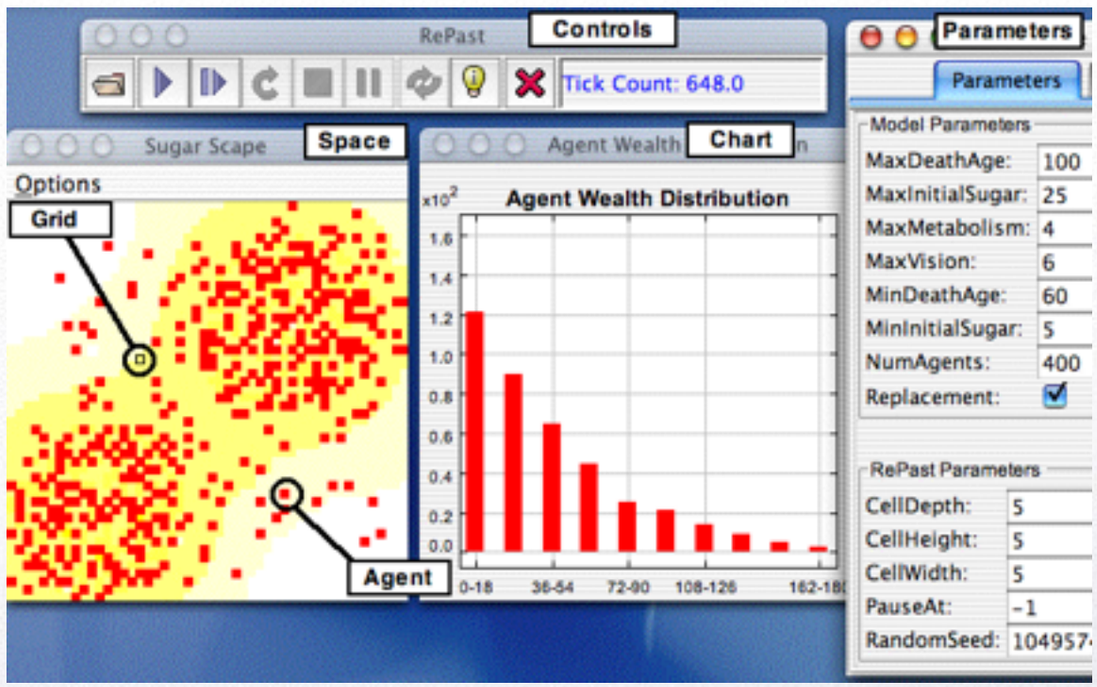




Before 2000

...larger, simulated societies with agent-based models





E.g., Is rule set X sufficient to generate social world Y.

Before 2000

In Artificial Societies: The Computer Simulation of Social Life (1995)

"The EOS project: Integrating two models of palaeolithic social change"

J. Doran, M. Palmer

Before 2000

Intelligent Systems: a Semiotic Perspective (1996)

"Symbolic Interactionist Modeling: The Coevolution of Symbols and Institutions" by Deborah Vakas Duong

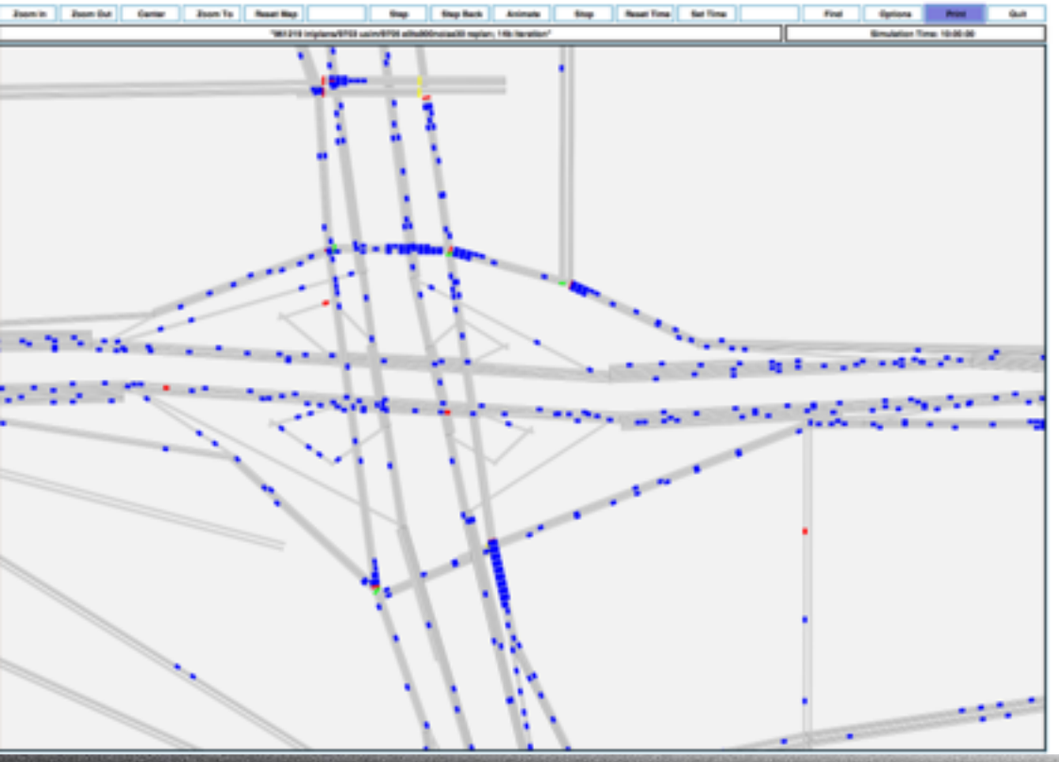
This paper presents a methodology for simulating social systems based on the sociological theory of symbolic interactionism. It is successful in emerging macrolevel institutions such as a role based division of labor and price from microlevel display and reading of signs. This paper presents the philosophy of this type of modeling and an agent based economics simulation which implements it.

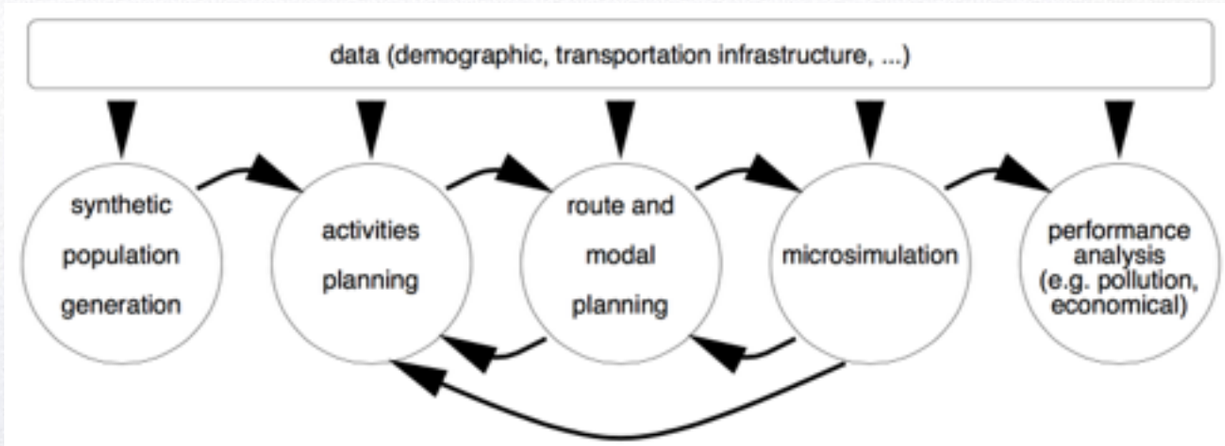
Before 2000

TRANSIMS for transportation planning

Kai Nagel, Richard J. Beckman, and Christopher L. Barrett

Los Alamos National Laboratory, Mail Stop M997,
Los Alamos NM 87545, U.S.A.





The TRANSIMS (TRansportation ANalysis and SIMulation System) project at the Los Alamos National Laboratory attempts to model all aspects of human behavior related to transportation in one consistent simulation framework. Input to TRANSIMS are transportation infrastructure data, demographic data, land-use data, and knowledge about human decision-making. Output data are any set of traditional or non-traditional measures of effectiveness (MOEs) that one could obtain from a second-by-second knowledge of the transportation system.

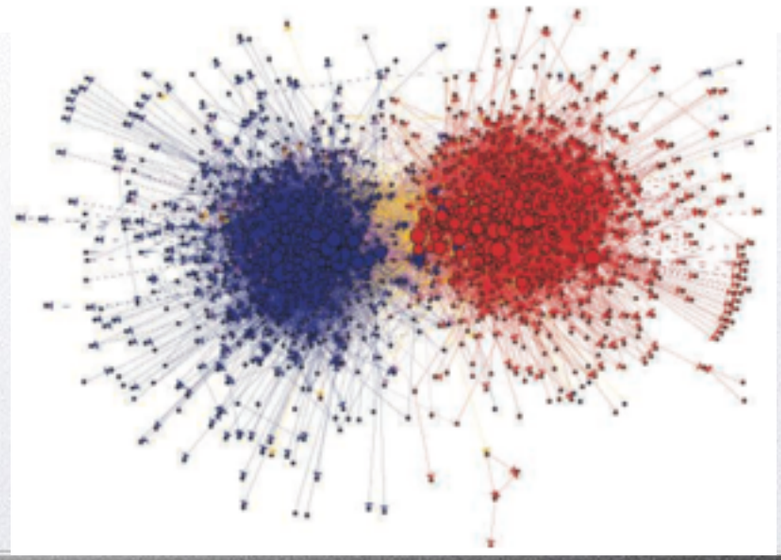
The key to TRANSIMS is a completely microscopic simulation of the travelers. Demographic data is disaggregated into synthetic individuals; these individuals make plans about their activities (work, sleep, eat, etc.) and how they intend to get their; finally, a realistic transportation microsimulation executes all plans simultaneously and thus computes the interactions between plans, for example resulting in congestion. Iterating between plans-making and plans-execution (the microsimulation) simulates day-by-day learning.

After 2000 (relabeled 2009)

SOCIAL SCIENCE

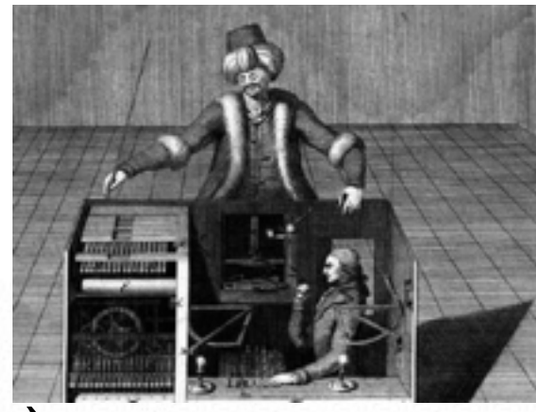
Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5]
Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3]
Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]



Social Data Revolution

High throughput:



1. Experiments (virtual labs.—e.g., MechTurk)
2. Observatories (sensors—e.g., Cell phones; RFID tags; Casual images/video; Quantified selves; Quantified crowds)
3. Archives (repositories—e.g., GoogleBooks, digitized government documents)
4. Surveys (crowdsourcing—e.g., active learning)

When does *Big Data* matter?

Rare (but consequential) Events:

- Viral events sparking collective attention
- Bridging network connections
- Novel behaviors or expressions
- Anything in the ***tail*** of the frequency distribution

When does *Big Data* matter?

MANAGEMENT SCIENCE
Vol. 62, No. 1, January 2016, pp. 180–196
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

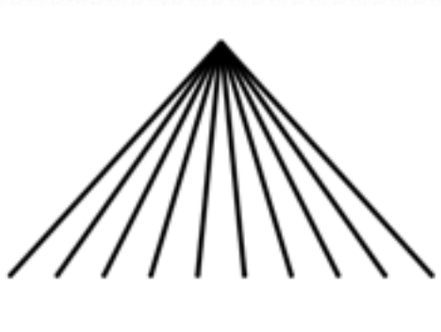
http://dx.doi.org/10.1287/mnsc.2015.2158

The Structural Virality of Online Diffusion

Sharad Goel, Ashton Anderson
Stanford University, Stanford, California, 94305 {scgoel@stanford.edu, ashton@cs.stanford.edu}
Jake Hofman, Duncan J. Watts
Microsoft Research, New York, New York 10016 {jmh@microsoft.com, duncan@microsoft.com}

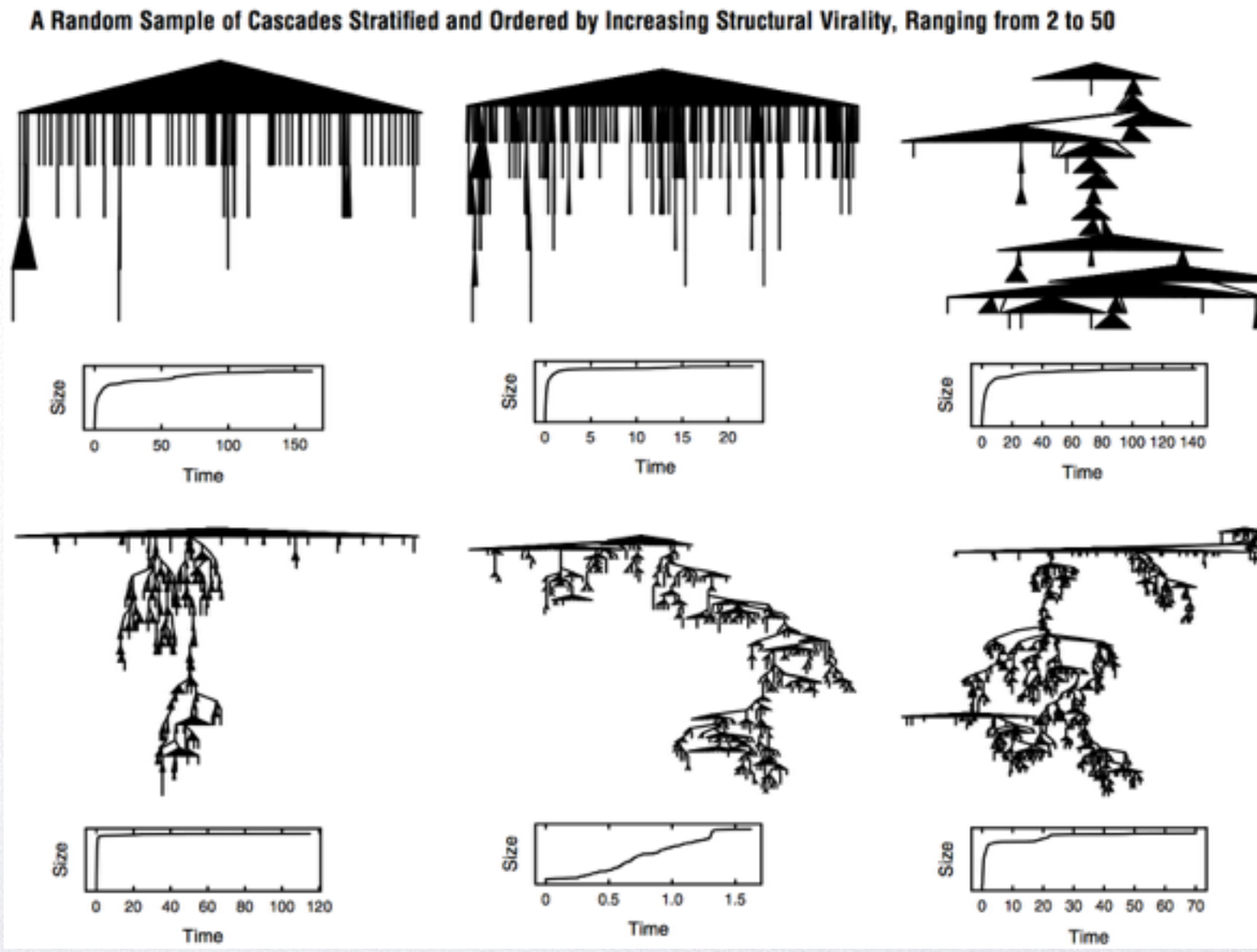
When does *Big Data* matter?

broadcast



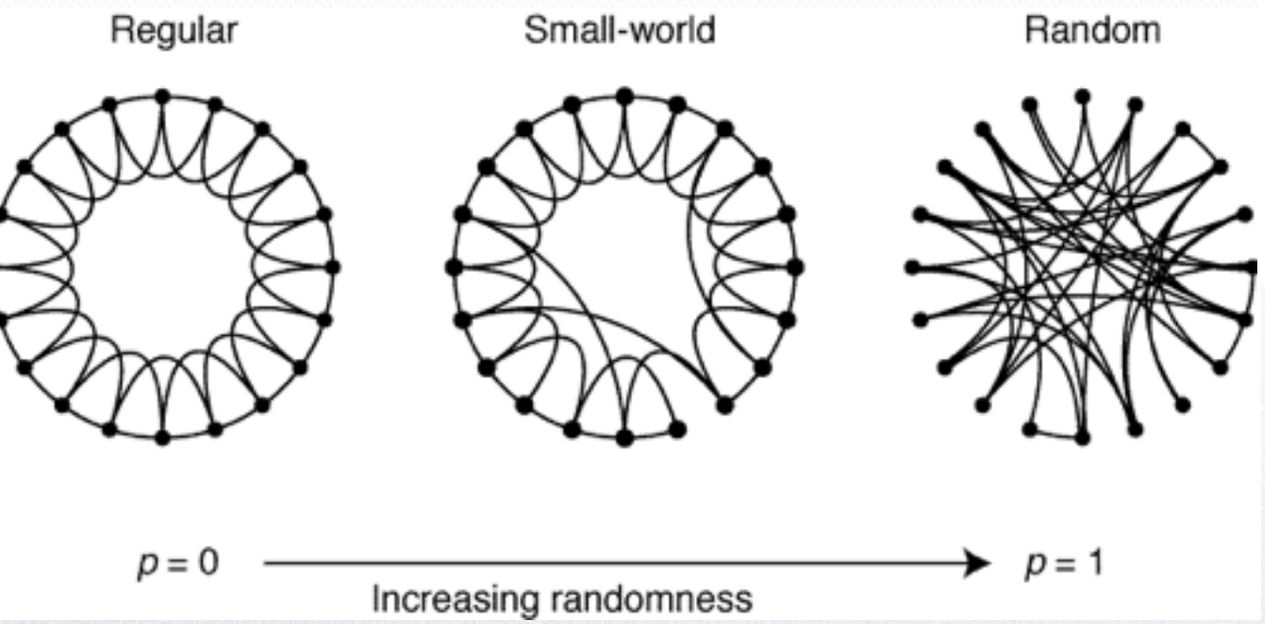
viral

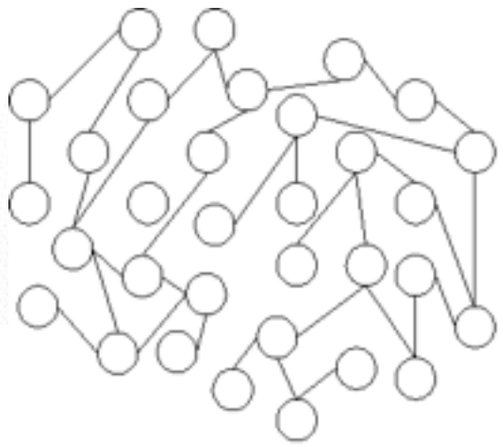
news stories, videos, images, and petitions



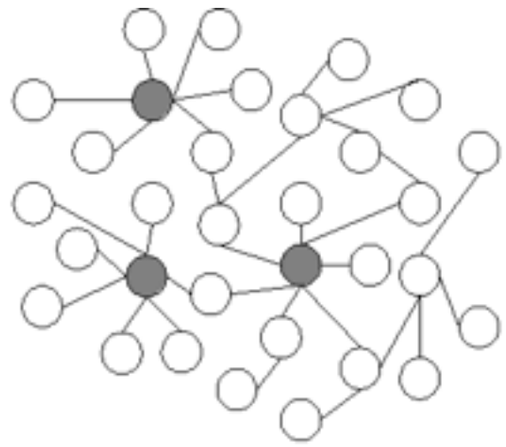
When does *Big Data* matter?

Small world network



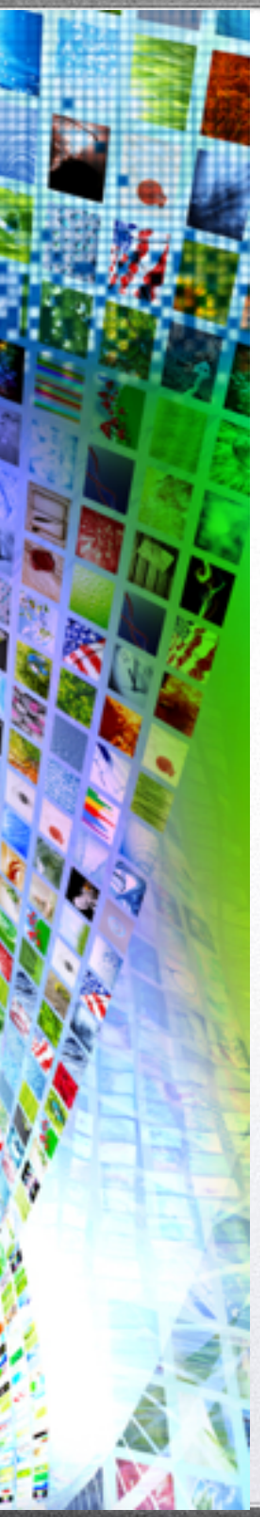


(a) Random network



(b) Scale-free network

Scale-free network



Social Data Revolution 2.0

Massive Datafication:

1. Natural Language Processing (Text, Audio ➡ Massive integration of admin. records)

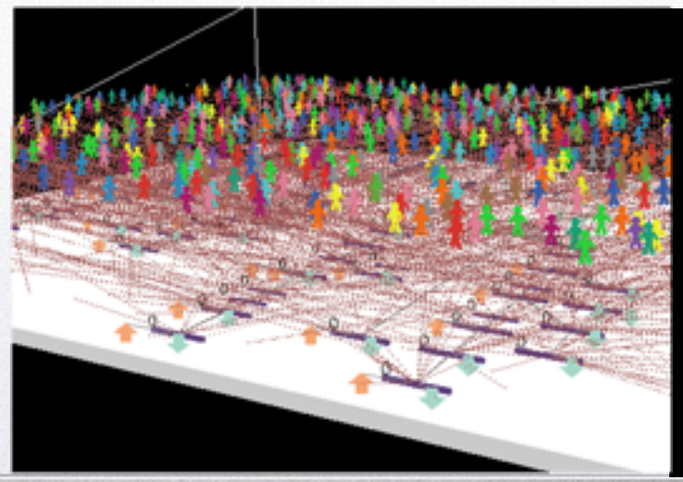
2. Machine Vision, Signal Processing (Images, video ➡ Views/Scenes; Social interaction and Network dynamics)

3. Machine Learning (General data tensors ➡ Reduce Dimensionality; Scale interpretive judgements; Intelligent sampling)

Simulated Data

High throughput Simulations:

1. Simple models (direct or agent-based; reproduce social phenomena)
2. Calibrated models (reproducing social phenomena at scale; with large substrates of existing data & multiple mechanisms)



Social Analytics Revolution

High Performance Computing:

1. Parallel High Performance Computing (Bayesian revolution, massive simulation, networks @ scale)
2. Cloud-based storage solutions (Computation over large, dispersed data; enclave computing; science as a service, distributed surveys and experiments)
3. *Artificial Intelligence (Algorithms for hypotheses generation; Hypothesis discovery then testing; Bots in the wild)

Social Analytics Revolution

High Dimensional Modeling:

1. LASSO ("Least Absolute Shrinkage and Selection Operator") & relatives

2. Random Forests (Computation over large, dispersed data; enclave computing; science as a service, distributed surveys and experiments)

3. Deep Learning (Algorithms for hypotheses generation; Hypothesis discovery then testing; Bots in the wild)

and new modes of statistical inference

Theoretical Entailments

Small data → Strong models

Big data → Weak models

Theoretical Entailments

Generative Standard for social scientific epistemology:

Don’t trust it if you can’t generate it

Theoretical Entailments

Social simulations:
plausible and sufficient theory

Data analysis:
discovering/testing mechanisms
predicting/generating data in question

Theoretical Entailments

Shift from causal inference to prediction

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0}\underbrace{(Y)}_{\text{prediction}} + \frac{\partial \pi}{\partial Y}\underbrace{\frac{\partial Y}{\partial X_0}}_{\text{causation}}$$

Shift from minimizing bias to balancing bias and variance

Theoretical Entailments

Combined and Competing Explanations

Granovetter's fascinating 1978 paper "Threshold models of collective behavior," published in *American Journal of Sociology*

"A theory of fads, fashion, custom, and cultural change as informational cascades," published in 1992 by Bikhchandani, Hirshleifer, Welch in *Journal of Political Economy*

Granovetter assumes implicitly that the order in which one observes one's peers does not change the information one infers from each of them, BHK assume quite explicitly that the order is of critical importance: what the focal actor i infers from a previous actor j taking action X depends on whether some third actor k has already taken the same action. BKH have a good reason for making this assumption—namely that if i knows that j knows what k has done, j's action means something different that if j is acting solely on his or her private information.

Synthetic Computational Social Science

- Compete and Combine Explanations
- Less intuitive; Less ascetically appealing
- Complexity artifact of world or data
- More relevant to policy, products, services
- end game…not subfield, social science

Opportunity

Phase-transition in explanatory
& predictive power of social science

Gains & Losses

Digital Archives & Observatories

Virtual Laboratories

Intelligent Surveys

Prediction vs. Inference…what kind?

…

Contributions

Risks

Social Science left behind

Machines race ahead (the parable of GRID computing)

Machinists race ahead (CSS in computer science…answers without questions)

Social Scientists left behind

Students left behind (we can't produce our own next generation)

Public left behind (only private data)

Competitive Landscape

Computational Powerhouses…

Stanford University - SS/CS Ds

Columbia University - Data Sci I

Cornell University - Info Sci D

MIT - Media Lab

Carnegie Mellon -

Opportunity…

UChicago distinction

Theoretical Ambition - Engage and *Transform* Canonical Questions within and between Disciplines

Methodological Diversity - Use Computation in a wide(r) variety of ways—not just Data Science but Algorithmic Complexity and Systems

Retrofitting Education

Reinventing Scholarship